Journal of Modern Applied Statistical Methods

Volume 12 | Issue 2 Article 23

11-1-2013

Ordered Logit Regression Modeling of the Self-Rated Health in Hawai'i, With Comparisons to the OLS Model

Hosik Min
University of South Alabama, Mobile, AL, hksmin@gmail.com

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 371-380.

ISSN 1538 − 9472

Ordered Logit Regression Modeling of the Self-Rated Health in Hawai‘i, With Comparisons to the OLS Model

Hosik Min
University of South Alabama
Mobile, AL

Despite the ordinal nature of Self-Rated Health (SRH) variable, logistic regression models or regression models have been used without adequate justification for these applications. It is shown that ordered-logit regression model is the appropriate statistical strategy to estimate SRH, whereas the Ordinary LeastSquares model leads to biased conclusions.

Keywords: Ordered logit regression, OLS, ordinal outcome, self-rated health, health status

Introduction

Self-Rated Health (SRH) has long been a major research topic in health-related research (Mossey & Shapiro, 1982; Idler & Angel, 1990; Miilunpalo, Vuori, Oja, Pasanen, & Urponen, 1997; Eriksson, Unden, & Elofsson, 2001). The main reasons for this are that SRH can be used as an individual's general health status and/or an indicator of his or her quality of life and that the research importance of SRH will continue to increase because of a growing interest in health and healthy living (McMurdo, 2000; Eriksson, Unden, & Elofsson, 2001). Given the increased life expectancy and the aging of the population (NCHS, 2007), suffering and death from various diseases have declined, while the topic of healthy living has received greater attention (Row & Kahn, 1987; Glasgow, 2004; Glasgow, Min, & Brown, 2013). Health or lack thereof includes not only physical factors such as limitations to daily life activities (ADL) but also mental indicators such as SRH. As health condition and/or status can impact an individual's well-being in positive or negative ways, it is an important topic in public health.

Dr. Min is assistant professor in the Department of Sociology, Anthropology, and Social Work. Email him at: hksmin@gmail.com

Here the focus will be on methodological aspects; that is, the appropriateness of the ordered logit model for SRH, by comparing the results obtained using this method with those from the OLS model. SRH has often been measured as an ordinal variable; for instance, it is measured as a 5-point scale in this study (1=Poor, 2=Fair, 3=Good, 4=Very Good, and 5=Excellent). The analytical approach to handling this type of variable, however, is often logit regression (Avanath & Kleinbaum, 1997; Manor, Matthew, & Power, 2000; Pohlmann & Leitner, 2003) or Ordinary Least Squares (OLS) model (Winship & Mare, 1984; Wardle & Steptoe, 2003). The use of logit regression model can be easily denied because the logit model cannot deal with a dependent variable with more than two categorical and ordered outcomes in an appropriate way. In other words, if the SRH is developed as a dichotomous variable—e.g., poor versus good—and then a logit model is employed to estimate the logit coefficients, the results would lead to the loss of important information about the dependent variable (Hamilton, 1992; Berry, 1993; Hamilton, 1995; Avanath & Kleinbaum, 1997; Pohlmann & Leitner, 2003). In addition, only small percentage of Hawai‘i adults were having poor SRH (only 3%) in this study. Moreover, other kinds of social, cultural, and socioeconomic factors differentiating people who have good, very good, and excellent SRH will not be estimated if we use logit model.

Therefore, the goals of this paper are to present the methodological problems by comparing OLS, which often used to estimate ordinal outcome, and ordered logit models and to offer an easily understandable comparison of two methods by examining the likelihood of having a higher SRH in Hawai‘i. Considering wide use of OLS model for the dependent variable with many categories in ordered measurement (Mekelvey & Zavonia, 1975; Avanath & Kleinbaum, 1997), examining the statistical assumptions and violations the OLS model causes with ordered logit model would provide us a meaningful insights for employing an appropriate statistical methodology. In addition, this is a particularly important and relevant concern, given the expected increase in interest in general health status, both physical and mental.

As was indicated (Hawkes, 1971; Reynolds, 1973; Mekelvey & Zavonia, 1975; O’Brien, 1982), analyzing an ordinal variable with an ordinal regression model could lead to incorrect conclusions by violating the assumptions of the ordinal regression model. The OLS model has several assumptions known as a best linear unbiased estimating method (BLUE) (Hamilton, 1992; Berry, 1993; Hamilton, 1995; Avanath & Kleinbaum, 1997; Menard, 2001). For instance, the OLS model expects the dependent variable as linear and continuous one; the OLS model assumes that the mean of errors of prediction in the population regression

function must be zero; and the variance of the error term is constant for all values of independent variables, homoscedasticity

If the dependent variable is ordinal, however, these assumptions in general are not met (Mekelvey & Zavonia, 1975; Fox, 1991; Hamilton, 1992; Berry, 1993; Hamilton, 1995; Avanath & Kleinbaum, 1997). First of all, the ordinal dependent variable is non-linear, the values are presented in 0 to 1 probability as in a logit regression model; a non-linear model must have a different error structure and the error term does not have constant variance. As McKelvey and Zavoina (1975) argued, the OLS model may, in some cases, have the undesirable effect of causing regression analysis to severely underestimate the relative impact of certain variables. Accordingly, the ordered logit model, instead OLS model is considered to be the most appropriate methods if the dependent variable is ordinal to estimate more accurately (Hawkes, 1971; Reynolds, 1973; Mekelvey & Zavonia, 1975; O'Brien, 1982; Avanath & Kleinbaum, 1997; Pohlmann & Leitner, 2003).

Consequently, the best-fitting and most appropriate statistical model for handling the ordinal outcome is an ordered or probit model. This study, however, will use and focus on an ordered logit model, because the results of these two methods are similar and the ordered logit model is more common and its results are easier to interpret (Long & Freese, 2003).

Data and Methods

As described above, to measure the overall assessment of respondents' health, self-rated health (SRH) is used as a dependent variable. SRH is measured by a five-point scale and thus has a categorical and ordered nature. The best-fitting statistical model for handling the ordered outcome is known as an ordered-logistic regression model, which will be used as an analytical model here.

Here is an explanation of the ordered logit regression model. For the sake of explanation, symbols rather than actual variable names will be used (Long & Freese, 2003). Posit that Y is an ordinal dependent variable with c categories, and $\Pr(Y \le j)$ denotes the probability that the response on Y falls in category j or below (i.e., in category 1, 2, …, or j). This is called a cumulative probability. It equals the sum of the probabilities in category j and below:

$$\Pr(Y \le j) = \Pr(Y = 1) + (\Pr(Y = 2) + ... \Pr(Y = j) \quad (1)$$

A "c-category Y-dependent variable" has c cumulative probabilities: $\Pr(Y \le 1)$, $\Pr(Y \le 2)$, …, $\Pr(Y \le c)$. The final cumulative probability uses the entire scale; as a consequence, therefore, $\Pr(Y \le c) = 1$. The order of forming the final cumulative probabilities reflects the ordering of the dependent variable scale, and those probabilities themselves satisfy:

$$\Pr(Y \le 1) \le \Pr(Y \le 2) \le ... \le \Pr(Y \le c) = 1 \tag{2}$$

In an ordered logit model, an underlying probability score for an observation of being in the i^{th} response category is estimated as a linear function of the independent variables and a set of cut points. The probability of observing response category i corresponds to the probability that the estimated linear function, plus random error, is within the range of the cut points estimated for that response.

$$\begin{aligned} &\Pr\left(\text{Response Category for the j}^{th}\ \text{Outcome} = i\right) = \\ &\Pr\left(k_{i-1} < b_1 X_{1j} + b_2 X_{2j} + ... + b_k X_{kj} + u_j \le k_i\right) \end{aligned} \tag{3}$$

It is necessary to estimate the coefficients $b_1, b_2, ..., b_k$ along with cut points $k_1, k_2, ..., k_{i-1}$ where i is the number of possible response categories of the dependent variable. The coefficients and cut points are estimated using maximum likelihood.

To do this, the data used in this paper were obtained from the 2005 Hawaii Health Survey (HHS). The HHS is a representative-sample survey based on household, administered as a telephone interview survey to adult residents in more than 6,000 households each year. The principle objective of the survey is to provide statewide estimates of population parameters that describe (1) the current health status of the population; (2) respondents' access to and utilization of health care; and (3) the distribution of the population by age, sex, and ethnicity (SMS Research & Marketing Services, Inc., 2006).

The ordered logit regression model is thus estimated for the Hawai'i residents that predict their SRH using other socio-demographic and locale characteristics that have been shown in the demographic literature to be associated with SRH (Mossey & Shapiro, 1982; Idler & Angel, 1990; Kennedy, Kawachi, Glass, & Prothrow-Stith, 1998; Kawachi, Kennedy, & Glass, 1999;

Eriksson, Unden, & Elofsson, 2001). The controlling variables pertain to age, sex, race/ethnicity, marital status, education, and residential location. Some are measured as dummy variables and others as interval.

The variables are as follows: 1) Age is measured in years from age 18 to 99; 2) Male is a dummy variable indicating whether the respondent is male; if yes, it is coded as 1; 3) Married is a dummy variable indicating whether s/he is married; if yes, it is coded as 1; 4) Hawaiian is a dummy variable indicating whether the respondent is Native Hawaiian; if yes, it is coded as 1; 5) Japanese is a dummy variable indicating whether s/he is Japanese American; if yes, it is coded as 1; 6) Filipino is a dummy variable indicating whether the respondent is Filipino American; if yes, it is coded as 1; and 7) Other is a dummy variable indicating whether s/he belongs to Other ethnic categories; if yes, it is coded as 1 (with White used as the reference group); 8) Education is measured as 6 categories from illiterate to 4 or more years of college education (1=Illiterate/Only Kindergarten; 2=Grade 1 to 8; 3=Grade 9-11; 4=Grade 12 or GED; 5=College, 1 to 3 years; 6=College, 4 years or more); 9) Big Island is a dummy variable indicating whether the respondent lives in Big Island; if yes, it is coded as 1; 10) Kauaʻi is a dummy variable indicating whether s/he lives in Kauaʻi; if yes, it is coded as 1; 11) Maui is a dummy variable indicating whether the respondent lives in Maui; if yes, it is coded as 1 (with Oʻahu used as reference variable).

Results of Ordered Logit Regression Versus OLS Analysis

Table 1 presents frequency distributions for all independent variables as well as the dependent one. The average score of SRH for Hawaiʻi residents was 3.57, which lies between good and very good. The average age was 47.6 years old among the adult population (age 18 and over). Half of them were male (49%). Six out of ten Hawai'i adults were married (60%). As for race/ethnicity, 21% were Native Hawaiian, 22% were Japanese American, 15% were Filipino, and 17% were Other. The average level of education was 4.86, or close to 1-3 years of college education. As for residence, 13% lived in Big Island, 5% lived in Kauaʻi, 12% lived in Maui, and the remaining 70% lived in Oʻahu.

Table 2 presents the results of the ordered-logistic regression and the OLS analysis for Hawaiʻi adults in 2005. The results show that overall model fit was significant for both models, and most coefficients in both models were significant. The older the respondent, the lower the SRH; if s/he was married, s/he was more likely to have a higher SRH; compared to white respondents, all other racial and

ethnic categories, such as Native Hawaiian, Japanese American, Filipino American, and Other, show a lower likelihood of having a higher SRH. Also, as expected, the more educated the respondent, the higher the SRH; a person living in Kauaʻi and Maui has a higher likelihood of having higher SRH compared to a person living in Oʻahu.

Table 1. Descriptive Statistics from the 2005 Hawaii Health Survey (n=898, 593, weighted)

Variable	Mean	Std. Dev.
Self-rated Health	3.57	1.04
Age	47.60	17.59
Male	0.49	0.50
Marital Status		
Married	0.60	0.49
Race/Ethnicity		
Hawaiian	0.21	0.41
Filipino	0.15	0.35
Japanese	0.22	0.42
Other	0.17	0.38
Socioeconomic Status		
Education	4.86	1.02
Residence Island		
Big Island	0.13	0.34
Kauaʻi	0.05	0.22
Maui	0.12	0.32

The results, however, indeed present the evidence of inappropriateness of using OLS model compared to the ordered logit model. The male variable provided important information regardless of whether an ordered logit model or OLS was used to deal with an ordinal dependent variable. A male was shown to have a higher likelihood of having a higher SRH compared to female counterparts in the ordered logit regression model, but not in the OLS. As previous studies have pointed out, using an OLS model for an ordinal-dependent variable indeed produces this inconsistent and biased result: It could be concluded that male did

not have any effect on SRH, which would be crucially misleading in the OLS model. In addition, all the values of the coefficients in the OLS model were severely underestimated compared to those of the ordered logit model, which lessened the effects of contributing factors on SRH.

Table 2. Comparison of the Analysis Results of Ordered Logit Regression and OLS from 2005 Hawaii Health Survey (n=898, 593, weighted)

	Ordered Logit Regression			OLS		
Variable	**b**	**z**		**b**	**t**	
Age	-0.026	-220.65	*	-0.014	-232.6	*
Male	0.013	3.38	**	-0.003	1.68	
Marital Status						
Married	0.143	35.35	*	0.083	38.11	*
Race/Ethnicity						
Hawaiian	-0.583	-98.76	*	-0.298	-94.64	*
Japanese	-0.59	-10.18	*	-0.297	-97.14	*
Filipino	-0.642	-98.69	*	-0.315	-90.29	*
Other	-0.406	-65.85	*	-0.207	-62.84	*
Socioeconomic Status						
Education		162.95	*	0.176	165.57	*
Residence Island						
Big Island		0.71		0.005	1.52	
Kaua‘i		3.71	*	0.032	6.74	*
Maui		11.52	*	0.031	9.38	*
	LR Chi^2	106,789.24		F	10,379.23	
	Pseudo-R^2	0.043	*	Adj. R^2	0.113	*

* p<.05; ** p<.001
Note: The values of cut points to ordered logit regression and the value of constants for OLS are not shown here.

Discussion

This paper deals with an appropriate use of statistical modeling that frequently occurs when modeling ordinal variables, Self-Rated Health, which is measured using a 5-point scale here. By comparing the results of ordered logit regression and OLS models, this study could illustrate the potential problems with using OLS in the analysis of ordinal SRH variables. While most of the conclusions from the OLS model were similar to those from the ordered logit regression model, significant differences do exist. Most of all, the insignificance of male in the OLS model could lead to incorrect conclusions regarding this variable. In fact, the significant and positive effect for male had on a respondent's SRH score was revealed when this study used the ordered logit model. Furthermore, the OLS model underestimated the effects of all coefficients.

Accordingly, this study appears to show that the use of an ordered logit regression model is statistically appropriate for the modeling of Self-Rated Heath, which has an ordinal characteristic, in Hawai'i's adult population. More specifically, the use of the ordered logit regression model could help avoid inconsistent and biased conclusions and their detrimental effects on public health policy.

Considering the fact that the importance of studying health status indicators such as SRH continues to rise, the use of an appropriate analytical strategy will be invaluable in the future.

References

Avanath, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology, 26,* 1323-1333.

Berry, W. (1993). *Understanding regression assumptions*. 1st ed. Sage Publications, Inc.

Eriksson, I., Unden, A., & Elofsson, S. (2001). Self-rated health. Comparisons between three different measures. Results from a population study. *International Journal of Epidemiology, 30,* 326-333.

Fox, J. (1991). *Regression diagnostics: an introduction*. 1st ed. Sage Publications, Inc.

Glasgow, N. (2004). Healthy aging in rural America. In (L. W. M. N. Glasgow, N. E. Johnson, Eds.) *Critical issues in rural health* (pp. 271-281). Ames, Iowa: Blackwell Publishing Professional.

Glasgow, N., Min, H., & Brown, D. (2013). Volunteerism of older in-migrants and long-term residents in rural retirement destinations. In N. Glasgow & E. H. Berry (Ed.), *Rural aging in 21st century America* (pp. 231-250). New York: Springer Publishing Company.

Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. 1st ed. Cengage Learning.

Hamilton, L. C. (1995). *Data analysis for social scientists*. 1st ed. Boston, MA: Duxbury Press.

Hawkes, R. K. (1971). The multivariate analysis of ordinal measures. *The American Journal of Sociology, 76*(5), 908-926.

Idler, E. L. & Angel, R. J. (1990). Self-rated health and mortality in the NHANES-1 epidemiologic follow-up study. *American Journal of Public Health, 80*, 446-452.

Kawachi, I., Kennedy, B. P., & Glass, R. (1999). Social capital and self-rated health: A contextual analysis. *American Journal of Public Health, 89*(8), 1187-1193.

Kennedy, B. P., Kawachi, I., Glass, R., & Prothrow-Stith, D. (1998). Income distribution, socioeconomic status, and self rated health in the United States: Multilevel analysis. *BMJ: British Medical Journal, 317*, 917-921.

Long, S. J., & Freese, J. (2003). *Regression models for categorical dependent variables using STATA*. A Stata Press Publication. STATA Corporation. College Station: TX.

Manor, O., Matthew, S., & Power, C. (2000). Dichotomous or categorical response? Analysing self-rated health and lifetime social class. *International Journal of Epidemiology, 29*, 149-157.

Mckelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

McMurdo, M. E. T. (2000). A healthy old age: Realistic or futile goal? *BMJ: British Medical Journal, 321*, 1149-1151.

Menard, S. (2001). *Applied logistic regression analysis*. 2nd ed. Sage Publications**,** Inc.

Miilunpalo, S., Vuori, I., Oja, P., Pasanen, M., & Urponen, H. (1997). Self-rated health status as a health measure: The predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. *Journal of Clinical Epidemiology, 50*(5), 517-528.

Mossey, J. M., & Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. *American Joumal of Public Health, 72,* 800-808.

National Center for Health Statistics. (2007). *Health, United States, 2007 with chartbook on trends in the health of Americans*. Hyattsville, MD.

O'Brien, R.M. (1982). Using rank-order measures to represent continuous variables. *Social Forces, 61*, 144-155.

Pohlmann, J. T., & Leitner, W.W. (2003). A comparison of ordinary least squares and logisitc regression. *Ohio Journal of Science*, 103(5), 118-125.

Reynolds, H. T. (1973). On "the multivariate analysis of ordinal measures". *American Journal of Sociology, 78*(6), 1513-1516.

Row, J. W., & Kahn, R. L. (1987). Human aging: Usual and successful. *Science, 273*, 143-149.

SMS Research & Marketing Services, Inc. (2006). *HHS update: Hawaii department of health, office of health status monitoring. Hawaii Health Survey, 2004, Procedure Manual*. Honolulu, HI.

Wardle, J., & Steptoe, A. (2003). Socioeconomic differences in attitudes and beliefs about healthy lifestyles. *Journal of Epidemiology & Community Health, 57*, 440-443.

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review, 49*, 512-525.